3-1-2008

An Improved SNR Estimator for Speech Enhancement

Yao Ren
Marquette University

Michael T. Johnson
Marquette University, michael.johnson@marquette.edu

Accepted version. Published as a part of the proceedings of the conference, *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008: ICASSP; Las Vegas, NV, March 31, 2008 - April 4, 2008*, 4901 - 4904. DOI.

An Improved SNR Estimator for Speech Enhancement

Yao Ren
Speech and Signal Processing Laboratory, Marquette University, Milwaukee, WI
Michael T. Johnson
Speech and Signal Processing Laboratory, Marquette University, Milwaukee, WI

Abstract: In this paper, we propose an MMSE a priori SNR estimator for speech enhancement. This estimator has similar benefits to the well-known decision-directed approach, but does not require an ad-hoc weighting factor to balance the past a priori SNR and current ML SNR estimate with smoothing across frames. Performance is evaluated in terms of estimation error and segmental SNR using the standard logSTSA speech enhancement method. Experimental results show that, in contrast with the decision-directed estimator and ML estimator, the proposed SNR estimator can help enhancement algorithms preserve more weak speech information and efficiently suppress musical noise.

Section 1.

Introduction

The Ephraim-Malah (logSTSA) filter[1] is a Minimum Mean Square Error (MMSE) estimator of the clean speech spectral amplitude for speech enhancement. One important factor in the logSTSA filter is the smoothing behavior of the decision-directed (D-D) *a priori* SNR estimator which has significant impact on reducing musical noise artifacts. This

IEEE International Conference on Acoustics, Speech and Signal Processing, 2008: ICASSP; Las Vegas, NV, March 31, 2008 - April 4, 2008, (2008): pg. 4901-4904. DOI.

estimator is the weighted sum of two terms, the SNR value from the previous frame and an ML SNR estimate from the current frame. Various aspects of the approach for SNR estimation have been investigated in previous work. Cappe[2] demonstrated this estimator reduces low-level musical noise by limiting the smallest allowable value of the *a priori* SNR. Recently, Erkelens et al.[3] suggested the insertion of a compensation factor to correct the bias caused by the decision-directed approach, Plapous and Marro[4] implemented a method to improve the estimator adaptation speed, and Hasan et al.[5] designed an adaptive scheme for updating the weighting factor. All of these approaches focus on the adaptation component or weighting factor of the SNR estimator, retaining the ML estimation approach. In this paper, we use a new approach and directly derive an MMSE estimator of the *a priori* SNR, which results in an expression that implicitly factors in information from previous frames.

Thus this estimator combines the information from both parts of the original D-D estimator in an MMSE sense, without requiring an experimentally pre-specified weighting factor.

In Section 2, we review the D-D *approach* of Ephraim Malah. Section 3 presents a derivation of the proposed MMSE *a priori* estimator. Results are presented and discussed in Section 4, with conclusions in Section 5.

Section 2.

A priori SNR estimation

The decision-directed *a priori* SNR estimator of Ephraim and Malah[6] is given by

$$\hat{\xi}_{k,D-D}(n) = \alpha \frac{\hat{A}_k^2(n-1)}{\lambda_d(k,n-1)} + (1-\alpha)P[\gamma_k(n)-1] \quad (1)$$

where $\xi_k(n)A_k(n)\lambda_d(k,n)$, and $\gamma_k(n)$ denote the *a priori* SNR, the spectral amplitude, the noise variance and the *a posteriori* SNR of the kth spectral component in the nth analysis frame, respectively. The P function is given by:

IEEE International Conference on Acoustics, Speech and Signal Processing, 2008: ICASSP; Las Vegas, NV, March 31, 2008 - April 4, 2008, (2008): pg. 4901-4904. DOI.

$$P[\gamma_k(n)-1] \triangleq \begin{cases} \gamma_k(n)-1 & \gamma_k(n)-1 \geq 0 \\ 0 & otherwise \end{cases} \quad (2)$$

This estimator is called a "decision-directed" type estimator, because it is updated based on the previous frame's amplitude estimate. As can be seen from the equation, the first term comes from the amplitude estimator of the previous frame while the second term is an ML estimate of ξ_k determined from the *a posteriori* SNR $\gamma_k(n)$.

The motivation for using an ML approach is that ML estimation can estimate an unknown parameter of a given PDF without any prior assumptions on the parameter. This estimator maximizes the joint conditional PDF of noisy spectral amplitude $\gamma_k(n)$ given clean signal variance $\lambda_x(k)$ and noise variance $\lambda_d(k)$.[6]

$$\hat{\lambda}_{x,ML}(k) = \arg\max_{\lambda_x(k)} p(Y_k(n)|\lambda_x(k),\lambda_d(k)) \quad (3)$$

This estimation results in the ML *a priori* SNR estimator:

$$\hat{\xi}_{k,ML} = \begin{cases} \frac{1}{L}\sum_{l=0}^{L-1}\gamma_k(n-l)-1 & if\ nonnegative \\ 0 & otherwise \end{cases} \quad (4)$$

which can be easily implemented using a recursive average as follows:

$$\overline{\gamma}_k(n) = a\overline{\gamma}_k(n-1) + (1-a)\frac{\gamma_k(n)}{b} \quad (5)$$

$$\hat{\xi}_k(n) = P[\overline{\gamma}_k(n)-1] \quad (6)$$

where a and b are pre-specified constants.

IEEE International Conference on Acoustics, Speech and Signal Processing, 2008: ICASSP; Las Vegas, NV, March 31, 2008 - April 4, 2008, (2008): pg. 4901-4904. DOI.

Analysis[2] has shown that the underlying characteristic for eliminating musical noise artifacts lies in the recursive calculation of the *a priori* SNR of (1): when γ_k stays below or equal to 0dB, the second term is zero and the *a priori* SNR becomes a smoothed version of the *a posteriori* SNR; whereas when γ_k is larger than 0dB, the second term dominates and the $\hat{\xi}_k$ estimate follows the γ_k estimate very closely. In conjunction with the log-STSA filter, this behavior leads to smoothly increasing noise attenuation in low-energy and speech absent segments of the signal.

Section 3.

Proposed Mmse Estimator

The weighting factor α in (1) provides a tradeoff between the *a priori* SNR from preceding frames and the current *a posteriori* SNR, smoothing out the overall SNR estimate trajectory. This factor is often set as 0.98,[6] based on experimental performance. Ideally this smoothing factor should be a variable that is small during the transient parts of the waveform to allow rapid adaptation and is large during steady speech segments.[7] Here we approach the problem of SNR estimation from an MMSE estimation perspective, leading to the elimination of the empirical weighting factor in favor of an estimator that directly incorporates previous frame information.

This new estimator is derived to minimize mean square error of the *a priori* SNR estimation. In,[6] the *a priori* SNR is defined as the ratio between the variances of the k^{th} spectral components of the speech and the noise

$$\xi_k = \frac{\lambda_x(k)}{\lambda_d(k)} = \frac{E\{|X_k|^2\}}{E\{|D_k|^2\}} \tag{7}$$

Similarly, we use the instantaneous values of speech and noise power to create an *a priori* SNR random variable z_k,

$$z_k = \frac{a_k^2}{d_k^2} \tag{8}$$

IEEE International Conference on Acoustics, Speech and Signal Processing, 2008: ICASSP; Las Vegas, NV, March 31, 2008 - April 4, 2008, (2008): pg. 4901-4904. DOI.

where a_k and d_k are the instantaneous spectral amplitudes of the speech and noise in the k^{th} frequency bin, respectively. An MMSE estimator of ξ_k is obtained from the conditional mean

$$\hat{\xi}_{k,MMSE} = E\{\frac{a_k^2}{d_k^2}|Y_k\}$$

(9)

Following the same assumptions as the traditional logSTSA filter, a^k and d^k are assumed Rayleigh distributed:

$$p(a_k) = \frac{2a_k}{\lambda_X(k)}\exp\{-\frac{a_k^2}{\lambda_X(k)}\}$$
$$p(d_k) = \frac{2d_k}{\lambda_D(k)}\exp\{-\frac{a_k^2}{\lambda_D(k)}\}$$

(10)(11)

For convenience of notation the *a priori* SNR is denoted as $zk=sk/nk=\Delta\ a2k/d2k$, so that sk and nk have exponential distributions

$$p(s_k) = \frac{2}{\lambda_X(k)}\exp\{-\frac{s_k}{\lambda_X(k)}\}$$
$$p(n_k) = \frac{2}{\lambda_D(k)}\exp\{-\frac{n_k}{\lambda_D(k)}\}$$

(12)(13)

This results in zk having the following distribution

$$f(z_k) = \int_0^\infty \frac{4\cdot n_k}{\lambda_x(k)\cdot\lambda_D(k)}\exp\{-\frac{n_k\cdot z_k}{\lambda_X(k)}-\frac{n_k}{\lambda_D(k)}\}dn_k$$
$$= -\frac{4\cdot\lambda_X(k)\cdot\lambda_D(k)}{(z_k\cdot\lambda_D(k)+\lambda_x(k))^2}$$

(14)

IEEE International Conference on Acoustics, Speech and Signal Processing, 2008: ICASSP; Las Vegas, NV, March 31, 2008 - April 4, 2008, (2008): pg. 4901-4904. DOI.

Under the assumed statistical model, $p(Y_k|z_k)$ is given by

$$p(Y_k|z_k) = \frac{1}{\pi \cdot (1+z_k) \cdot \lambda_D(k)} \cdot \exp\{-\frac{|Y_k|^2}{(1+z_k) \cdot \lambda_D(k)}\} \tag{15}$$

and the conditional mean $E\{z_k|Y_k\}$ is then given by

$$E\{z_k|Y_k\} = \frac{\int_0^\infty z_k \cdot p(Y_k|z_k) \cdot p(z_k) dz_k}{\int_0^\infty p(Y_k|z_k) \cdot p(z_k) dz_k}$$

$$= \frac{\displaystyle\int_0^\infty \frac{z_k}{1+z_k} \cdot \frac{1}{\left(z_k + \frac{\lambda_X(k)}{\lambda_D(k)}\right)^2} \cdot \exp\left\{-\frac{1}{(1+z_k)} \cdot \frac{|Y_k|^2}{\lambda_D(k)}\right\} dz_k}{\displaystyle\int_0^\infty \frac{1}{1+z_k} \cdot \frac{1}{\left(z_k + \frac{\lambda_X(k)}{\lambda_D(k)}\right)^2} \cdot \exp\left\{-\frac{1}{(1+z_k)} \cdot \frac{|Y_k|^2}{\lambda_D(k)}\right\} dz_k}$$

$$\triangleq f\left(\frac{\lambda_X(k)}{\lambda_D(k)}, \frac{|Y_k|^2}{\lambda_D(k)}\right) \tag{16}$$

Note that the expression for this final solution incorporates the previous amplitude estimate $\lambda_X(k)$ and thus can be thought of as "decision-directed" in the same sense as the traditional method of equation (1). This new function is evaluated by using numerical integration.

IEEE International Conference on Acoustics, Speech and Signal Processing, 2008: ICASSP; Las Vegas, NV, March 31, 2008 - April 4, 2008, (2008): pg. 4901-4904. DOI.

Section 4.

Experimental Results

To evaluate the performance of the new estimator, logSTSA enhancements are performed over 10 speech utterances taken from the TIMIT database[8]. A frame size of 32 ms with 75% overlap is used. Three different experimental runs are implemented

1. logSTSA filter with ML estimator from (5) and (6).
2. logSTSA filter with D-D estimator from (1).
3. logSTSA filter with proposed estimator from (16).

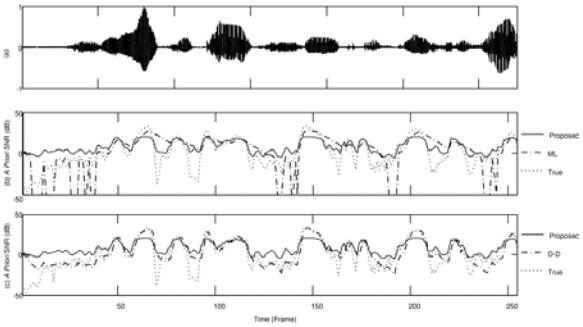

Fig. 1. A priori SNR estimated by 3 different methods in logSTSA filter.

The logSTSA filter itself is implemented using

$$\hat{A}_k = \frac{\hat{\xi}}{1+\hat{\xi}} \exp\left\{\frac{1}{2}\int_{v_k}^{\infty} \frac{e^{-t}}{t}\,dt\right\} R_k \quad (17)$$

The D-D weighting factor is taken as α=0.98 and the parameters for equation (5) and (6) are a=0.725, b=2. The a priori SNR is limited from −50dB to 50dB. White noise is added to each utterance at an Segmental SNR (SSNR) level of −10, −5, 0, +5, +10 dB. The noise spectrum is estimated by averaging the first 3 frames of each noisy utterance.

Evaluation of the method was done by comparing SNR estimation accuracy, objectively measuring quality of the enhanced signal through SSNR improvement, and by subjectively comparing spectrogram results.

IEEE International Conference on Acoustics, Speech and Signal Processing, 2008: ICASSP; Las Vegas, NV, March 31, 2008 - April 4, 2008, (2008): pg. 4901-4904. DOI.

An estimation comparison of the aforementioned three estimators is presented in Fig. 1. For the example plots in this figure, the frequency index bin *k* is 17, representing a center frequency of 562.5 Hz. The plots show the estimated *a priori* SNR (dB) and true *a priori* SNR in this specific frequency bin across the frames of a 0dB SSNR noisy utterance. Compared with the ML estimator, the proposed method avoids sudden drops and updates the estimate more smoothly. Qualitatively, this results in suppressing the musical noise often associated with the ML estimator. In high SNR regions, the proposed estimator matches the true value with smaller delay than the D-D estimator.

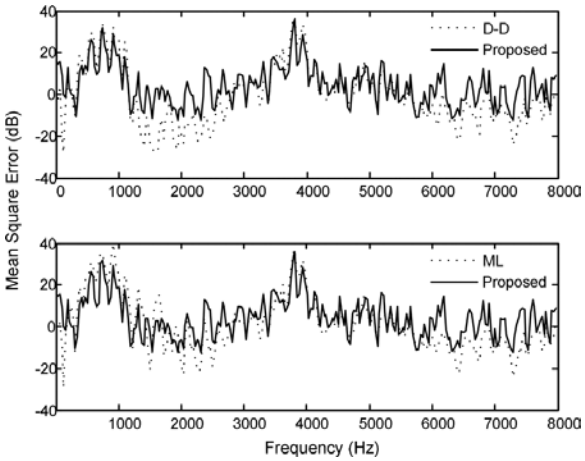

Fig. 2. Mean square errors of three estimation methods.

An example of mean square errors of the three estimation methods are shown in Fig. 2. This plot shows the averaged results across a 20 frame segment of the same utterance as in Fig. 1. Within the typical *speech* frequency range, 1000–5000 Hz, the proposed estimator has lower estimation error than both ML and D-D approaches. The quantitative results are shown in Table 1, including averaged results of MSE, median squared error (Med), standard derivation (Std) and interquartile range (IQR, 75%~25%). The proposed estimator has the lowest value for all four measures, which indicates this algorithm makes fewer estimation errors and is more reliable and robust.

Table 1. Estimation error.

	MSE (dB)	Std	Median	IQR
ML	168.51	230.89	366.97	25.76
D-D	110.60	162.23	160.03	10.03
Proposed	55.23	99.58	14.70	8.08

IEEE International Conference on Acoustics, Speech and Signal Processing, 2008: ICASSP; Las Vegas, NV, March 31, 2008 - April 4, 2008, (2008): pg. 4901-4904. DOI.

Objective evaluation results of logSTSA enhancement methods with three *a priori* SNR estimators are shown in Fig. 3. The averaged SSNR improvement from 10 utterances show *that* the proposed *method* has about 0.7 dB higher SSNR improvement than the original D-D estimator in the logSTSA filter. The higher SSNR results of ML are at the cost of introducing more musical noise.

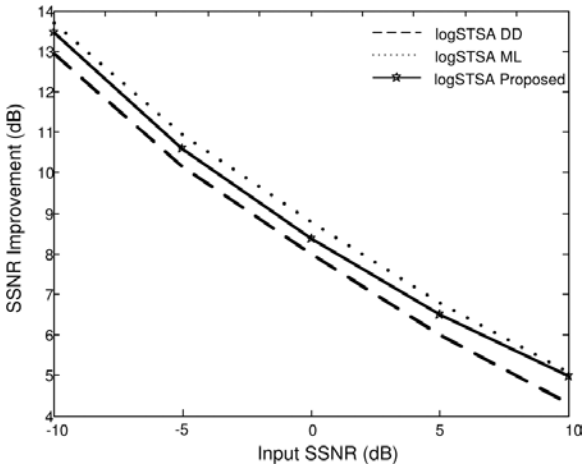

Fig. 3. SSNR evaluation of logSTSA filter with three a priori SNR estimators.

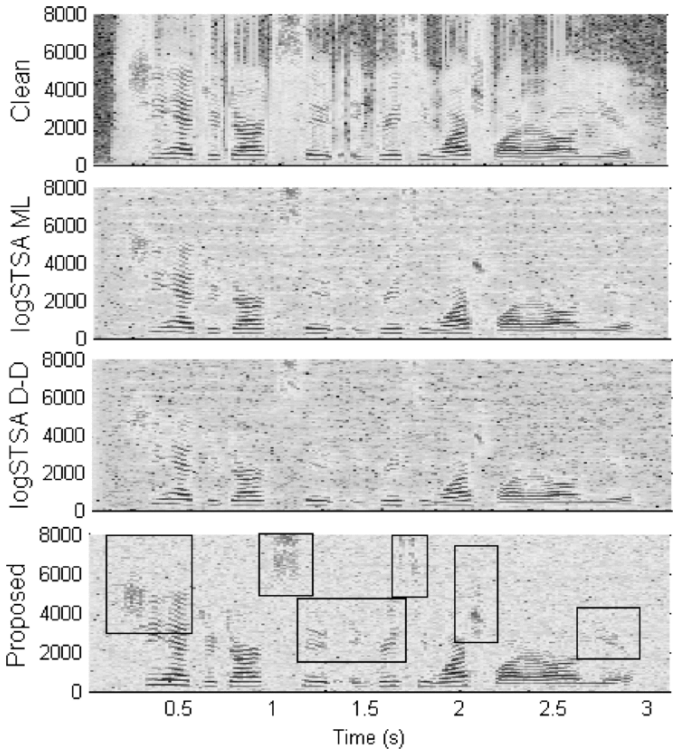

Fig. 4. Spectrograms of enhanced utterance by logSTSA filter with three a priori SNR estimators.

IEEE International Conference on Acoustics, Speech and Signal Processing, 2008: ICASSP; Las Vegas, NV, March 31, 2008 - April 4, 2008, (2008): pg. 4901-4904. DOI.

From the example spectrum in Fig. 4 we see that although improvement in SSNR is achieved by the ML estimator, a "musical noise" effect, consisting of small, isolated peaks in the spectrum, is introduced. Like the D-D estimator, the proposed estimator can suppress this artifact, due to the implicit smoothing action. Additionally, it can be seen that the proposed estimator helps preserve weak speech segment information more than D-D estimator, as shown in the highlighted rectangular areas, which also matches the estimation error results in Fig. 2.

Section 5.

Conclusion

A new *a priori* SNR estimator for speech enhancement is introduced in this paper. Unlike previous approaches to SNR estimation, this estimator is derived in the MMSE sense. The solution shows that this estimator implicitly incorporates the smoothing behavior of the original D-D estimator. Comparative results have shown that use of the proposed estimator in a logSTSA speech enhancement filter can effectively reduce noise as well as help preserve weak speech segment information.

References

[1]Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoustic. Speech Signal Process.*, vol. ASSP-33, 1985.

[2]O. Cappe, "Elimination of the musical noise phenonmenon with Ephraim and Malah noise suppressor," *IEEE Trans. Acoustic. Speech Signal Process.*, vol. 2, pp. 345-349, 1994.

[3]J. Erkelens, J. Jensen, and R. Heusdens, "A data-driven approach to optimizing spectral speech enhancement methods for various error criteria," *Speech Commun*, vol. in press, 2007.

[4]C. Plapous and C. Marro, "Improved Signal-to-Noise Ratio Estimation for Speech Enhancement," *IEEE Trans. on Audio, Speech, and Language Processing, 2006.*

[5]M. Hasan, S. Salahuddin, and M. Khan, "A modified a priori SNR for speech enhancement using spectral subtraction rules," *IEEE Signal Process. Lett.*, vol. 11, pp. 450-453, 2004.

[6]Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square short-time spectral amplitude estimator," *IEEE Trans. Acoustic. Speech Signal Process.*, vol. 32, pp. 1109-1121, 1984.

[7]P. C. Loizou, *Speech enhancement: theory and practice*, CRC press, 2007.

[8]J. Garofolo, L. Lamel, and W. Fisher, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM," *NISTIR* 4930, Feb, 1993.

IEEE International Conference on Acoustics, Speech and Signal Processing, 2008: ICASSP; Las Vegas, NV, March 31, 2008 - April 4, 2008, (2008): pg. 4901-4904. DOI.